DOCUMENT RESUME

ED 118 617 TM 005 105

AUTHOR Nicolich, Mark J.
TITLE Longitudinal Data Analysis with Pictures, Regression and Principal Components.
PUB DATE [75]
NOTE 19p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Data Analysis; Graphs; *Longitudinal Studies; Multiple Regression Analysis; *Statistical Analysis
IDENTIFIERS Change Point Analysis; Principal Components Analysis

ABSTRACT

Several statistical techniques that can be used to ameliorate the difficulties inherent in the data analysis of longitudinal studies are presented. The first step in longitudinal data analysis is graphing. This permits visual inspection of the data, and with educated viewing can yield insights into the nature of the underlying mechanisms. The next level of sophistication is to apply regression analysis and change point analysis to the curves obtained from the graphical analysis. It is usually the case in longitudinal studies that the exact form of the curve is not known prior to the experimentation. The graphing of the data is useful in suggesting different mathematical models to apply to the curves. The results of the regression analysis will help determine the uniformity of the process across subjects. The next step is to use the form of the fitted equation to determine significant points on the curve. The shape of the curve will suggest change points in the subjects' behavior with respect to the dependent variable. In certain cases where problems arise, the use of principal components is called for. Practical advantages are that they explain the original curve best and will likely point to any existing major differences, and they occur mathematically and do not depend on the experimenter's ability to form a regression curve or pick important change points. When used in conjunction with each other, these techniques form a powerful package for analyzing longitudinal data. (RC)

Longitudinal Data Analysis with Pictures, Regression and Principal Components

Mark J. Nicolich, Ph.D.
Statistician, Mathematics Department
Rider College, Trenton, New Jersey

In a longitudinal study, when several variables are investigated, the data analysis is often perplexing. The independent variable is invariably time, but the appropriate choice of the dependent variable and method of analysis is not always clear. There are several statistical techniques that can be used to ameliorate these difficulties, and they range from simple to apply and interpret to conceptually complex. The techniques are graphing, regression, change point analysis and principal components.

There are additional, high-powered, statistical techniques that can be used to analyze longitudinal data: cannonical correlation, multivariate analysis of variance, multivariate time series, and factor analysis with a wide choice of rotation procedures. While these techniques are applicable, their results are often difficult to interpret. When dealing with longitudinal data and subjects who are in a period of rapid development a major difficulty is that the response variables are varying with time, and not all responses vary at the same rate. This difficulty is further complicated by subject to subject variability. The beginning of a developmental trend is often not at the same chronological age for all

subjects; in addition the period to full development will vary from subject to subject. These forms of variability make analysis difficult. This paper presents suggestions for ameliorating these problems of longitudinal data analysis, where small samples are used.

The first step in an analysis of longitudinal data is to graph the data. This permits visual inspection of the data, and with educated viewing can yield insights into the nature of the underlying mechanisms. In almost all types of data analysis, the visual inspection of data is important. With longitudinal data the picture of the data is almost always rewarding.

A series of graphs should be prepared for each subject. For each subject a separate graph is made for each dependent variable of interest. All graphs should have the same time scale to make interpretations simpler. The graphs are then laid out in rectangle with similar measures lined up in one direction, and subjects lined up in the other direction. The "grid of graphs" is now open for inspection.

First look for patterns from subject to subject on a particular dependent measure. Is the shape of the curve the same for each subject? Are some displaced along the time axis? Are some elongated or heightened relative to the others? Some subjects may have developed early or late and not have full curves, so the remaining portion will have to be added mentally. If all curves are almost the same then there is little or no subject to subject variability on the given measure, and the phenomenon can be investigated using all subjects. If there is a displacement or distortion of the curve with respect to some subjects, determine what

distinguishes these subjects from the others. Is it only one point that is astray, did the subject have "a bad day" during that observation, was the subject "learning from that kid next door"? These reasons might account for subject to subject variability.

Next look for patterns in the measures from subject to subject. If a subject is relatively advanced in one measure, does it also hold for other measures? Does the pattern of differences coincide with the nature of the subject?

Patterns within subjects and patterns from subject to subject should emerge. If patterns do not show, then, possibly, the instruments used do not measure what they are intended to (they lack validity), or the underlying hypotheses need reorganizing, or the patterns are subtle and need more sophisticated analysis.

As an example, data from a one year longitudinal study was analyzed. Five female subjects, ranging in age from 14 to 19 months at the start of the study, were observed approximately monthly for a 40 minute period for the duration of the study. (Nicolich, 1975). Of the measurements made in the original study, consideration will be made of two linguistic variables (type imitation rate, and number of word types) and one non-linguistic variable (level of symbolic play). Figures 1 through 8 represent graphs derived from the data.

Figure 1 demonstrates several pattern changes for a single response variable for the five subjects. Tracy exhibits a shift, in that her pattern begins relatively late. Mira has her "steps" compressed and probably went through the levels between 15 and 16 months of age, and the steps were missed. Shanti has a relatively elongated pattern. All five

subjects have a similar pattern, but slight changes or individualities are noted.

Figure 2 graphs the type imitation rate for all five subjects on the same graph. It is difficult to make comparisons or conclusions from this graph. Five separate graphs, laid out side by side, yield a better picture, and promote interpretation. See Figures 4 through 8. The time axis is the same for the first four subjects but changed for the last in order to give a larger graph. Note it is Tracy, the late starter, who again has her pattern shifted to the right, this is also evident from Figure 2. Again the pattern of the five subjects is similar; a right skewed, unimodal, platykurtic distribution (a mound trailing off to the right). Mira's pattern is most different, and again is compressed. Shanti has an elongated pattern again. The beginning of Meri's curve is missing since the study began after the start of her development on this variable. The 18 month observation for Janis is out of pattern; notes made by the researcher at the time of the visit indicate that she was visited four hours later than usual and was tired and listless. This is an indication of the ability of the technique to highlight inconsistencies in the data.

Figure 3 is another example of similarity in subjects on yet another measure (broken into three sub measures). No sophisticated statistical techniques are required to show that, for all subjects studied, the spontaneous word types exceed the other categories after the first sessions. Other conclusions that can be made from the graph are that the number of spontaneous word types increase during the period under study while the other categories are relatively constant, the spontaneous-imitated category

is less frequent than the imitated and that at a time near the 5th session there is a sharp increase in the number of spontaneous word types.

The graphical technique can be used as preliminary analysis to more sophisticated techniques, to get a feel for the data and to detect patterns. It can also be a terminal analysis in some cases when the results are strong and clear cut.

The next level of sophistication is to apply regression analysis and change point analysis to the curves obtained from the graphical analysis. It is usually the case in longitudinal studies that the exact form of the curve is not known prior to the experimentation. The graphing of the data is useful in suggesting different mathematical models to apply to the curves. In most cases several different curves should be fit to the data to find a curve which is uniformly acceptable for the data from each subject. If there are any subjects that behave differently from the majority, then either analyze them separately or remove them from further analysis. They should receive special attention to determine why they are different; are they unique to themselves or do they form another phenomena?

The results of the regression analysis will help determine the uniformity of the process across subjects. If the significance level of the fitted curve is similar for all subjects, then it can be concluded that the phenomenon is uniform among the subjects. In the example, a curve of the form

$$y = Ax^{B}e^{Cx}$$

was found to fit the curves of the five subjects depicted in Figures 4 through 8. In the equation y is the type imitation rate, x is the age in months, e is the base of the natural logarithim and A, B and C are the

parameters to be estimated in the equation. The significance levels were .002, .02; .04, .02, and .06 for the regressions fitting Figures 4 through 8. The fit was uniform and good, thus it was concluded that the equation adequately modeled the phenomenon in these subjects. It can then be concluded that these five subjects all behave similarly with respect to time and type imitation rate.

The fitting of the regression curves is to determine uniformity of subject response. The next step is to use the form of the fitted equation to determine significant points on the curve. The shape of the curve will suggest change points in the subjects behavior with respect to the dependent variable. The change points might be onset of the phenomena, termination of the phenomena, peak value, time when a change in behavior is observed, or any other measure suggested by theory or curve shape. These change points are then investigated as additional, theoretical, data points. They can be considered to be error free, in that they are derived from a fitted model and not from direct observation. They are really error free only insofar as the fitted model is error free.

In the example being used, the maximum type imitation rate was considered to be of theoretical importance. The maximum, predicted from the fitted model, was found for each fitted curve by use of the differential calculus; the maximum is plotted on the graphs. There is close agreement of theoretical and actual maxima in four of the five curves; again it is Mira who doesn't follow as well. The maximum tends to be near the transition from play level 3 to play level 4. It also is near a large increase in the number of single word imitations (See Figures). These observations are the type that can be made from this kind of analysis.

The practical significance and usefulness of the observations depend on the data at hand. In the example the relationship of the maximum to the single word types was useful in explaining a linguistic phenomenon.

In addition to a visual inspection and interpretation of the change points, any type of data analysis can be used with the change points as data. A t test could be used to see if two subgroups of students (males, and females for example) differ in their change points. If the subjects were chosen according to an experimental design layout, the change points could be analyzed according to the design configuration, and the analysis would proceed as with any experimental design program. Alternatively the change points might be used as a dependent variable in a regression with the independent variable some measure on the subject such as age at the start of the study. When doing this type of analysis care must be taken. The usual caveats of analysis of variance and regression must be observed, and in addition, if several change points per curve are to be analyzed, remember they are not independent of each other and results from separate analyses will be correlated.

If the curves from the graphing are unwieldy, or do not lend themselves to regression analysis, or the regression results do not yield adequate change points, or if further analysis of important points is desired but the interdependency is a problem, then the use of principal components is called for. The full technique, with a complete example, is described by Church (1966).

Principal components analysis would use as input data measurements made on the dependent variable at chosen time points (the independent variable). The subjects would have to have been measured at the same point

in time. The time would be either the subjects age time or calendar time depending on what is used as the independent variable. If the dependent variable was not measured at the same time for each subject, the value of the dependent variable at the required time could be approximated by interpolation between adjacent time values.

Data for each subject would be a column of values of the dependent variable, each value taken at a determined point in time. The principal component analysis yields a set of linear combinations of these data points which "explain" the cause of the variability. A linear combination is a weighted average of the dependent variables taken over time. Each linear combination has a different weighting combination. The "explanation" is the accounting for the variable not having the same value throughout the time period. The weightings often have direct interpretations and have the same interpretation as the factors in factor analysis. An important fact is that the components are independent of each other.

Usually a subset of the principal components (3 to 5 of them) are sufficient to "explain" the majority of the variability. The weightings of the principal components can be applied to each subjects dependent variable measurements to come up with a set of individual, independent scores. These scores can be treated as change points, except they may have a more complex interpretation than that of a maximum, minimum, etc. These scores can then be used in the t test, analysis of variance or regression as previously explained. Their major statistical advantage of principal component scores over change points is that they are independent of each other and a separate analysis can be performed for each score type without regard for the correlation of score types. Practical advantages are that they are "best at explaining" the original curve and will likely

point to major differences if they exist, and they occur mathematically and do not depend on the experimenters ability to form a regression curve or pick important change points.

No example is presented for this technique because the data used in the examples did not require this level of sophistication. The inspection of change points was sufficient to indicate the underlying phenomena. Discussions and examples of principal components analysis can be found in Cooley and Lohnes (1971), Morrison (1967) and Van de Geer (1971).

The series of techniques described form a powerful package for analyzing longitudinal data. When used in conjunction with each other they can indicate new approaches and ideas concerning the data, as well as verifying predetermined hypotheses.

References

Church, Alonzo Jr., Analysis of data when the response is a curve. Technometrics, May 1966, 8(2), 229-246.

Cooley, W. W., & Lohnes, P. R., Multivariate Data Analysis. New York: Wiley and Sons, 1971.

Morrison, D. F., Multivariate Statistical Methods. New York: McGraw-Hill, 1967.

Nicolich, L. , A Longitudinal Study of Representational Play in Relation to Spontaneous Imitation and Development of Multiword Utterances. Doctoral dissertation, Rutgers University, 1975.

Van de Geer, J., Introduction to Multivariate Analysis for the Social Sciences. San Francisco: W. H. Freeman and Company, 1971.

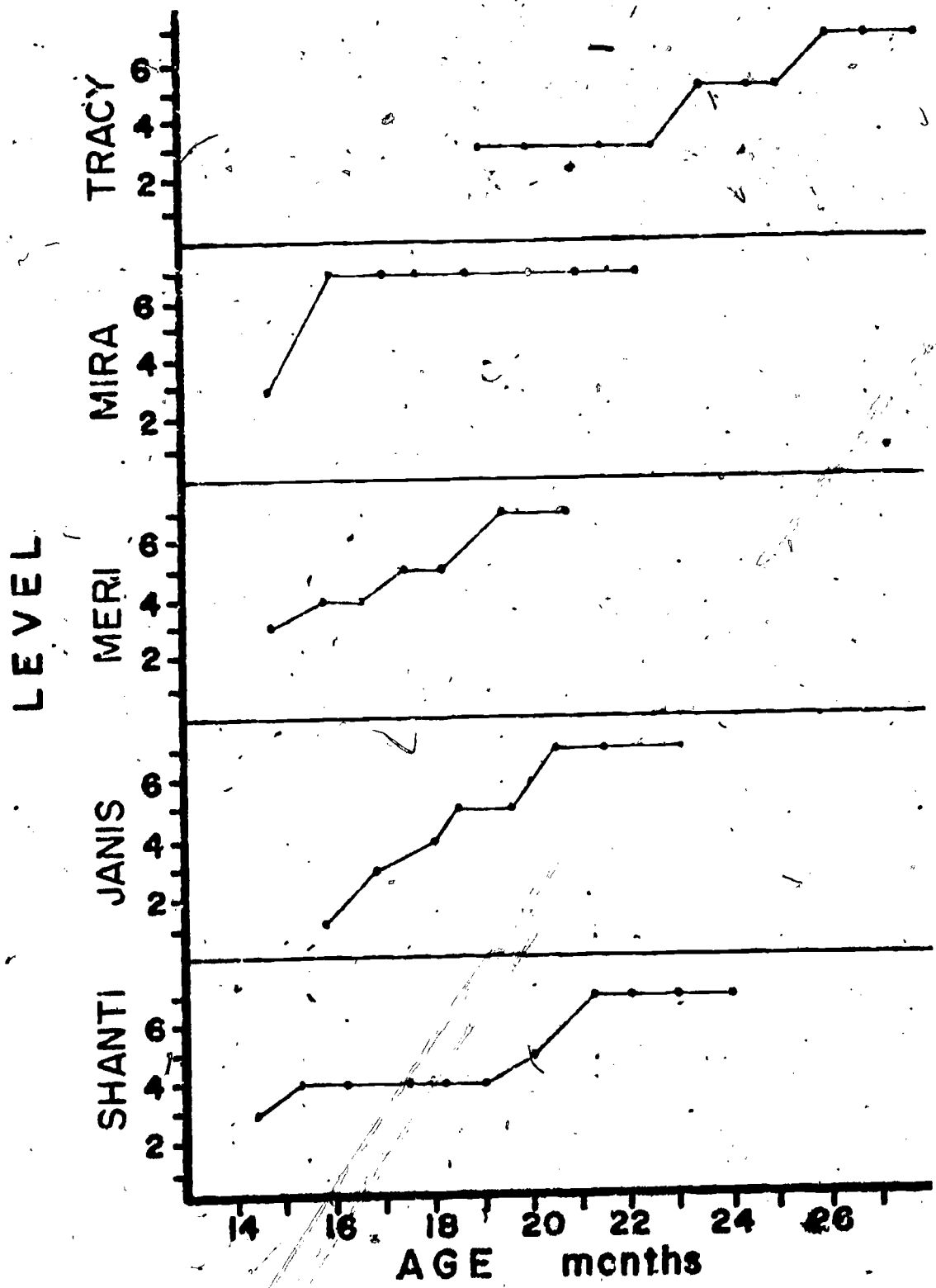


Figure 1. Age at attainment of symbolic play levels.

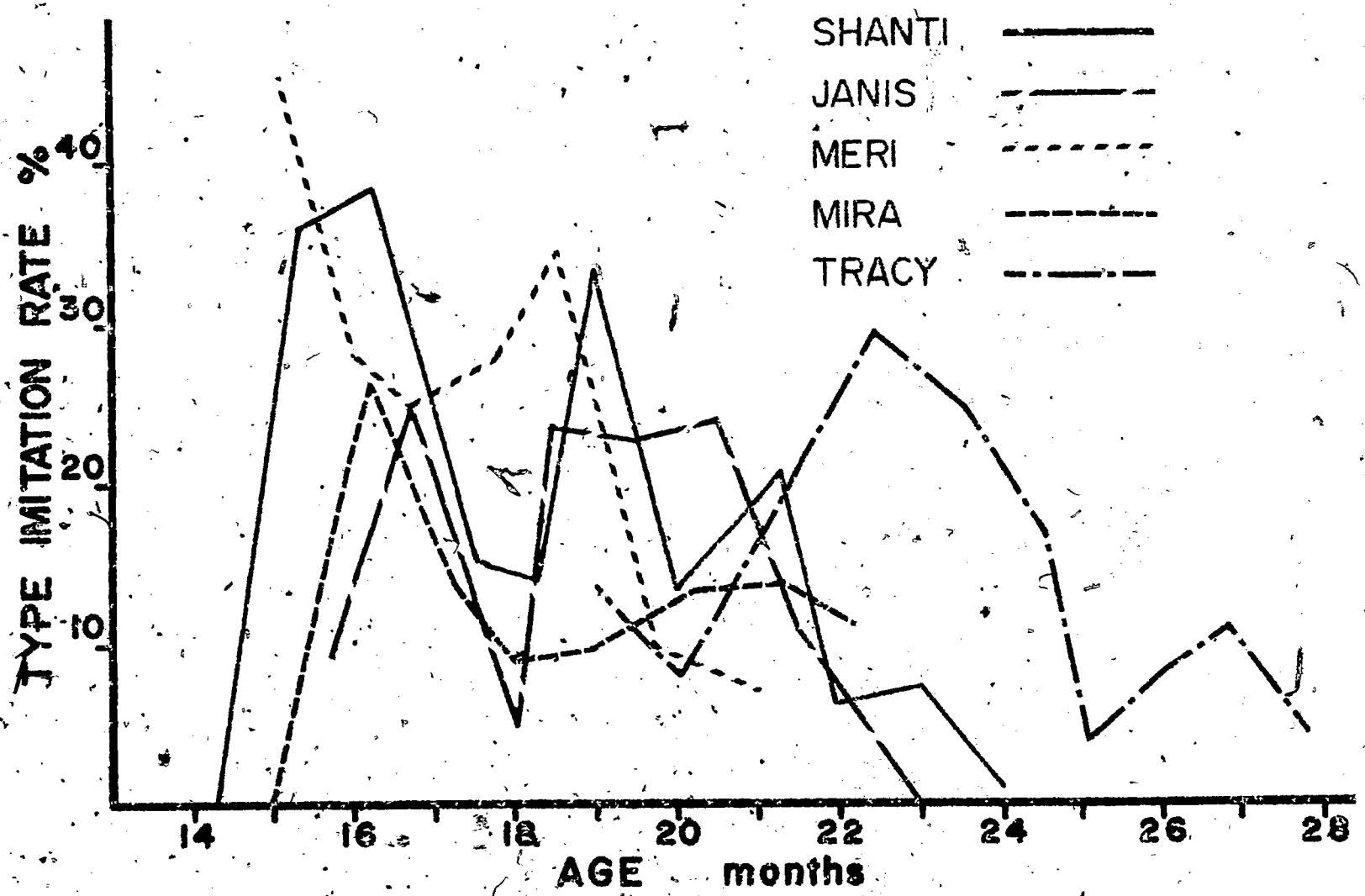


Figure 2. Proportion of word types imitated only.

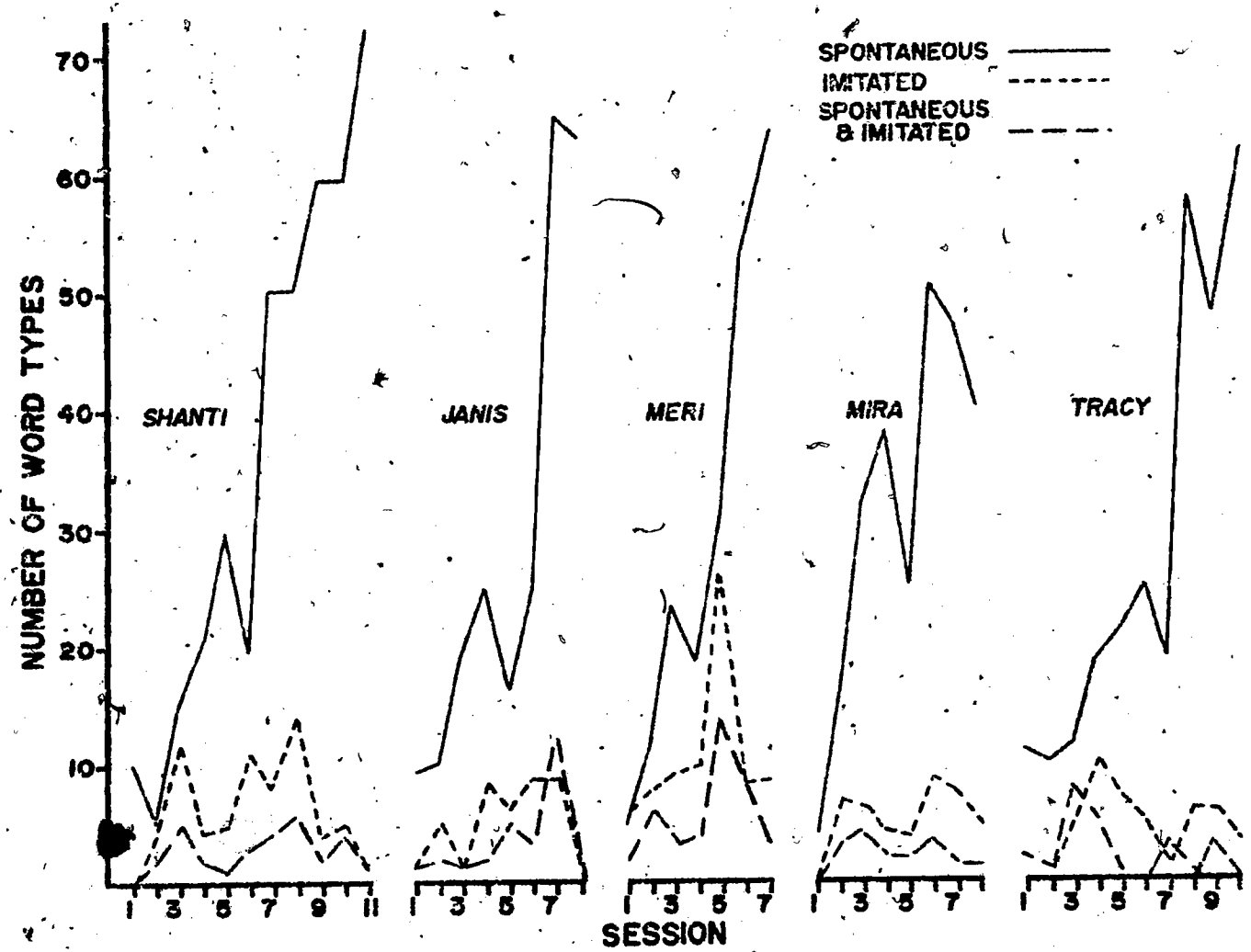
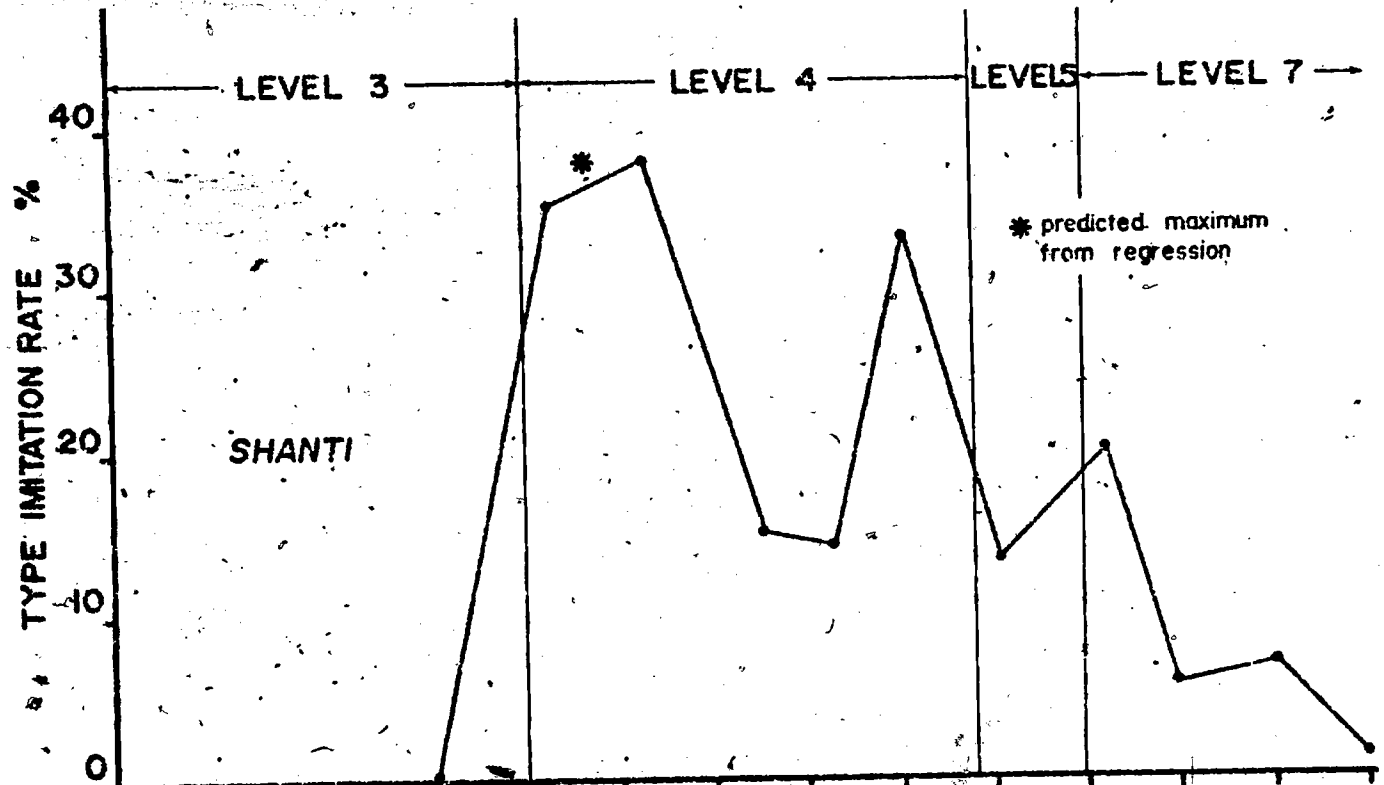


Figure 3. Number of single word types classified as imitated, spontaneous, and both imitated and spontaneous.



SESSION	1	2	3	4	5	6	7	8	9	10	11
ACTION-JUDGMENT TYPES	0	1	5	6	4	5	6	8	7	7	5
SINGLE TYPES	10	10	27	21	32	28	57	63	58	62	70
MULTI-WORD TYPES	0	0	0	1	4	0	18	95	109	120	134

Figure 4. Summary information: Shanti.

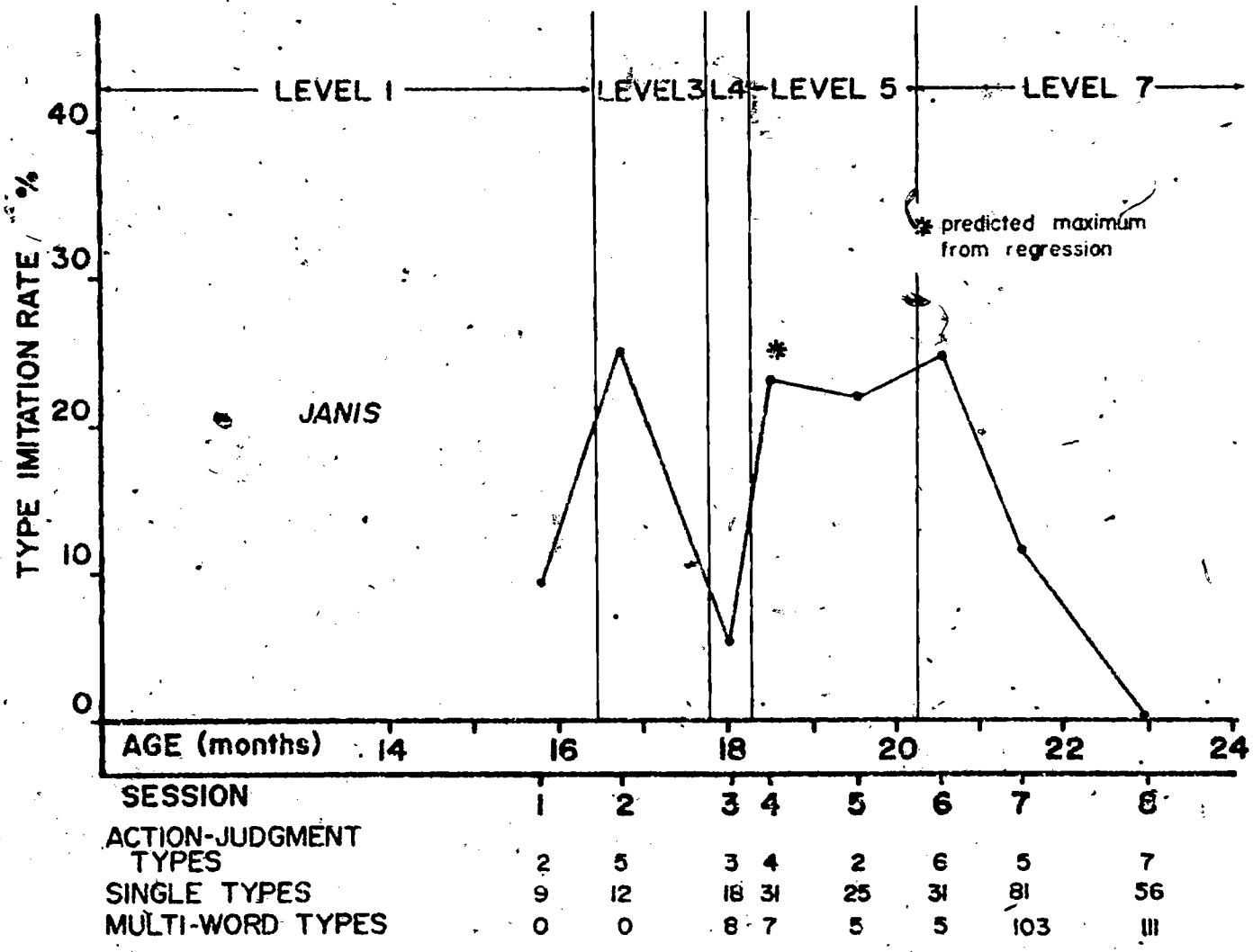


Figure 5. Summary information: Janis.

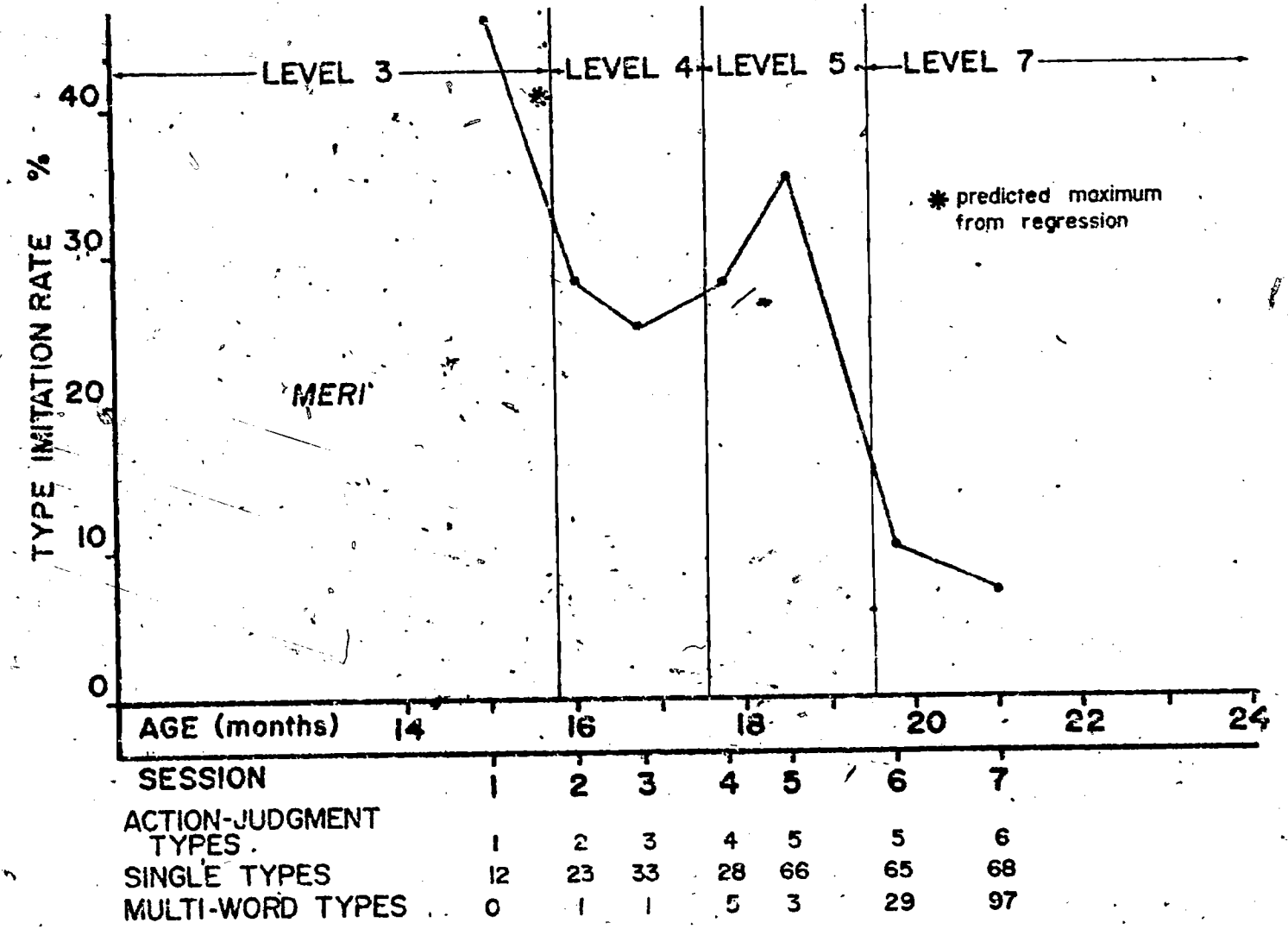


Figure 6. Summary information: Meri.

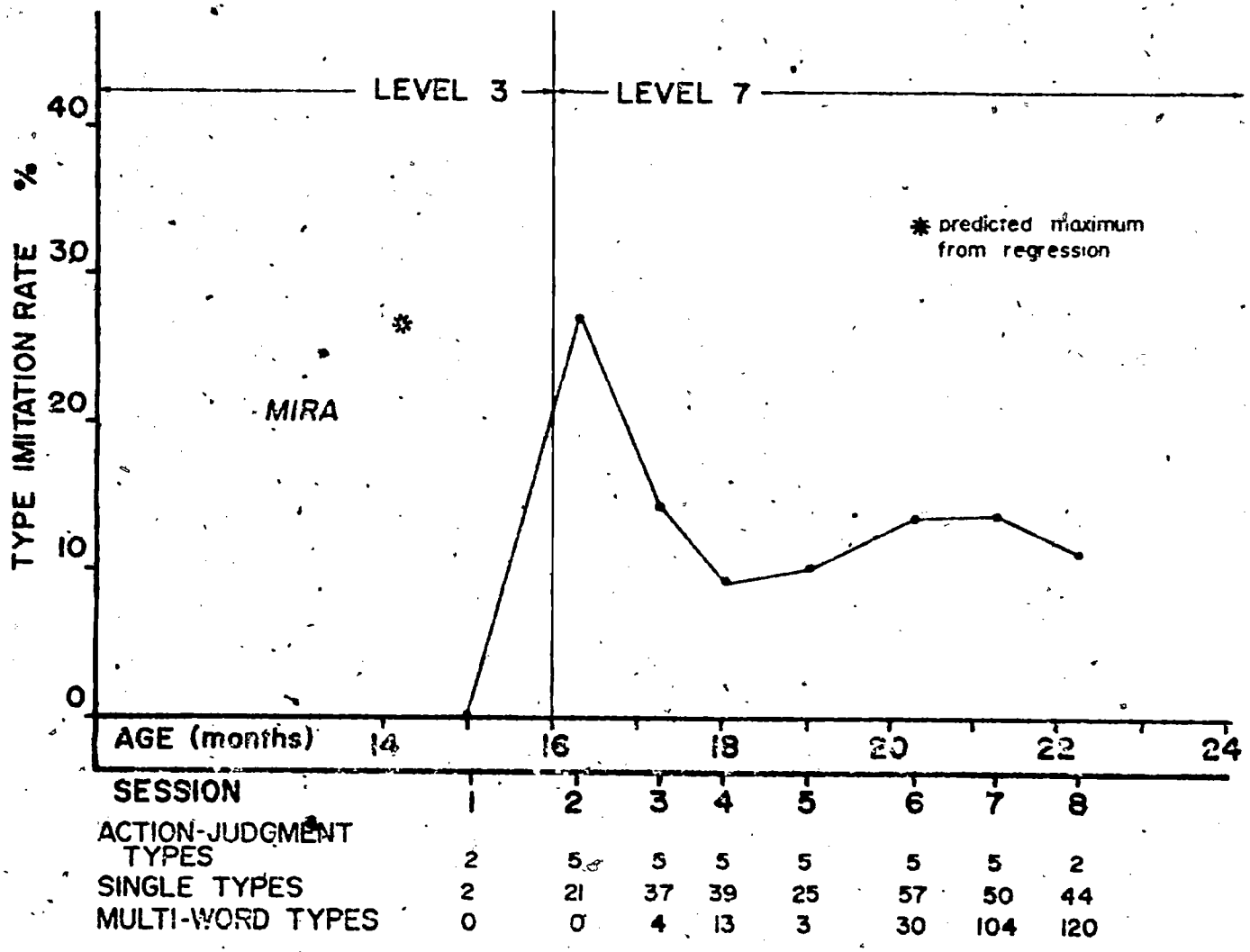


Figure 7. Summary information: Mira.

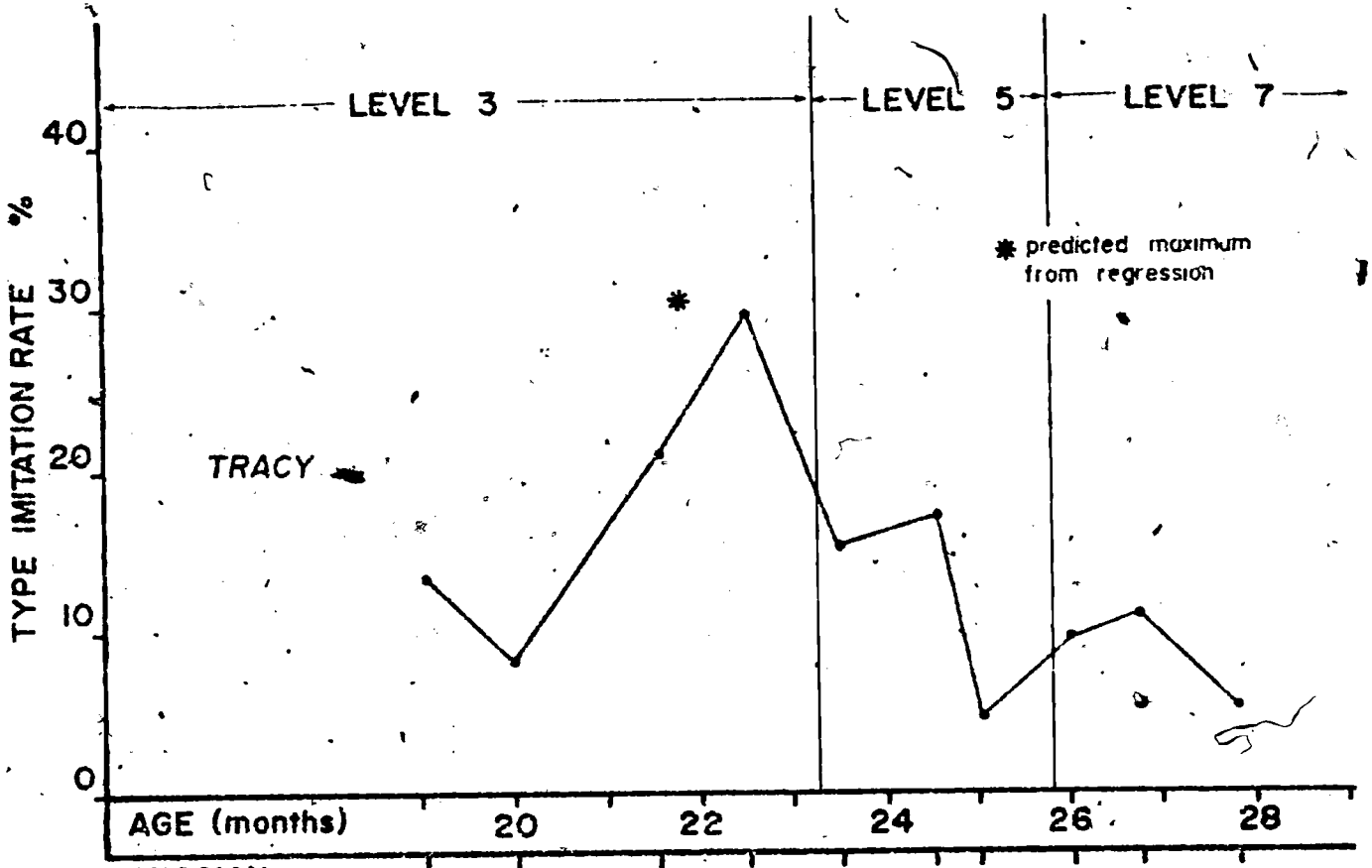


Figure 8. Summary information: Tracy.